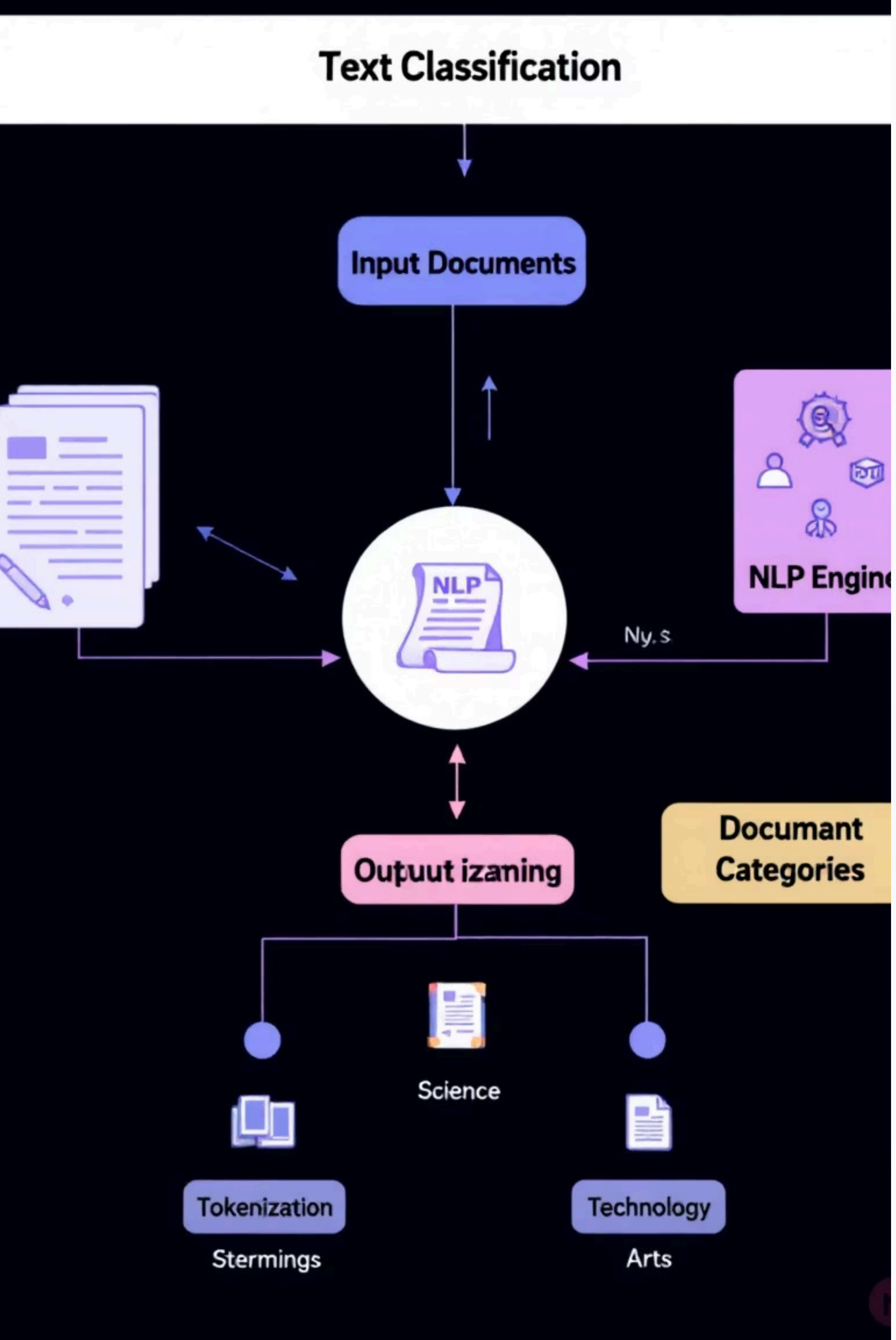


News Topic Classification using NLP

An NLP downstream task focused on classifying short news texts into four categories using Logistic Regression.

Project Objective

The Challenge

Automatically categorize news articles into relevant topics with high accuracy.

Create a complete machine learning pipeline from preprocessing to deployment.

- Classify news texts into four distinct categories
- Apply NLP techniques to textual data
- Build a user-friendly interface for real-time classification

Tools & Technologies

Language

Python was used as the primary programming language for its extensive ML libraries.

Libraries

scikit-learn, pandas, nltk, and gradio provided the necessary tools for ML and UI.

Platforms

Google Colab for model training and VS Code for deployment and refinement.

Dataset Overview

Dataset Composition

Labeled news titles and descriptions across four categories.

World, Sports, Business, and Science & Tech classifications.

Preprocessing Steps

Text Cleaning

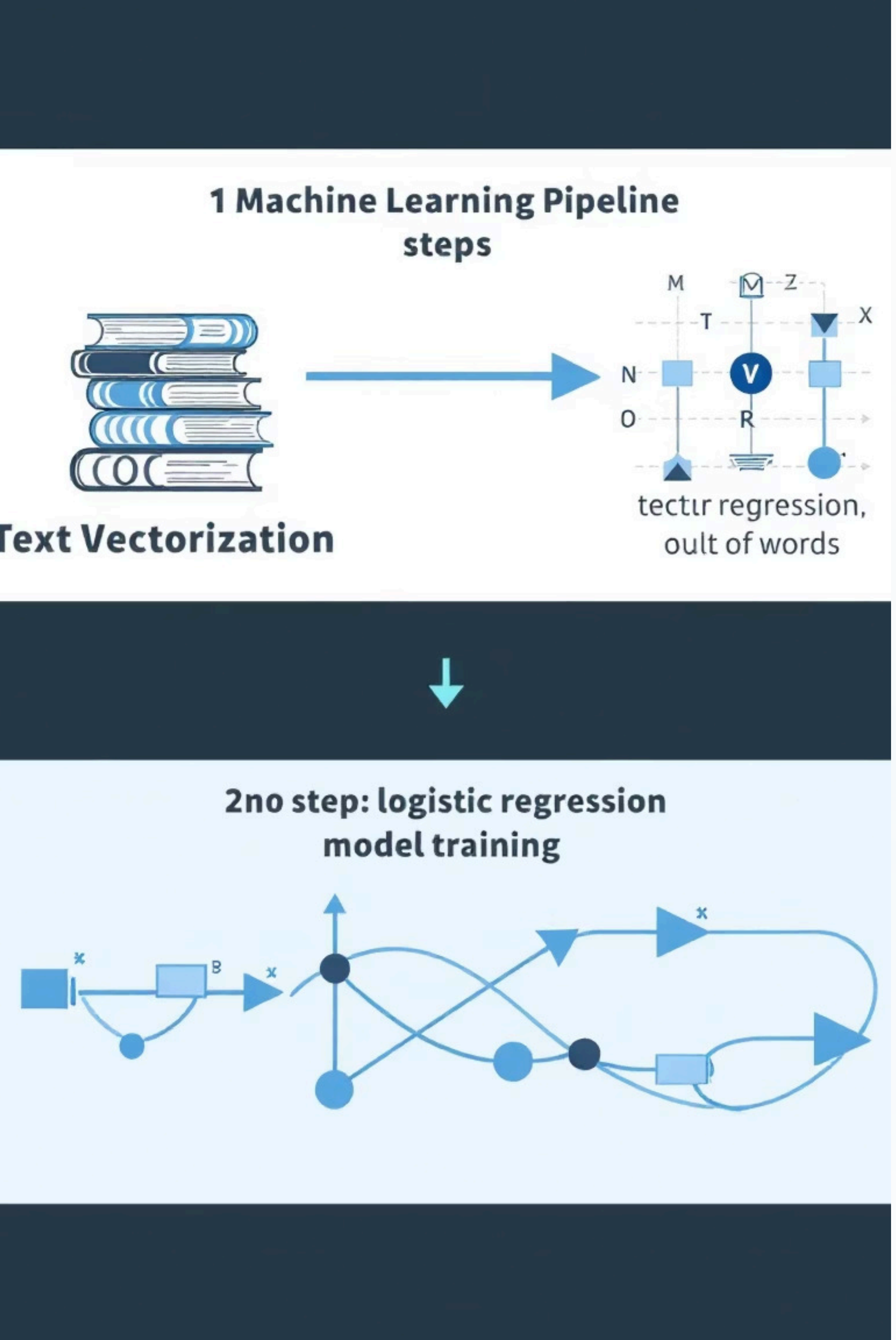
Regular expressions removed special characters and normalized text.

Stopword Removal

NLTK stopwords eliminated common words with low information value.

Text Normalization

Converted text to lowercase and removed extra whitespace.


1 Machine Learning Pipeline steps
M
M
Z
T
X
N
V
O
R
ext Vectorization
tectır regression,
oult of words
2no step: logistic regression model training
x
B
x
x

Model & Methodology

Text Vectorization

Applied **TF-IDF** to convert text into numerical features.

Captured word importance across the document corpus.

Data Splitting

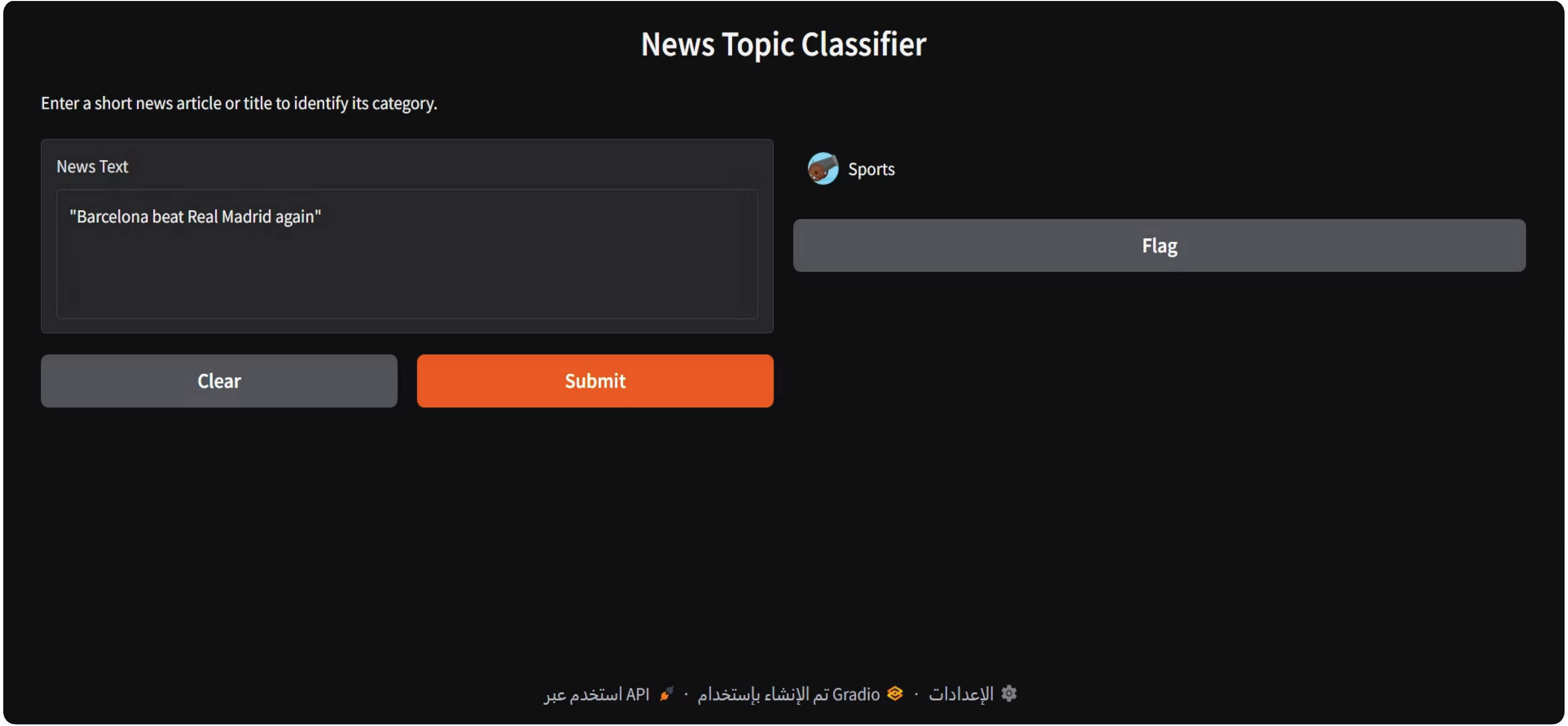
80% training data and 20% testing data.

Stratified split maintained class distribution.

Model Training

Logistic Regression with optimized hyperparameters.

Multi-class classification with 'one-vs-rest' approach.



News Topic Classifier

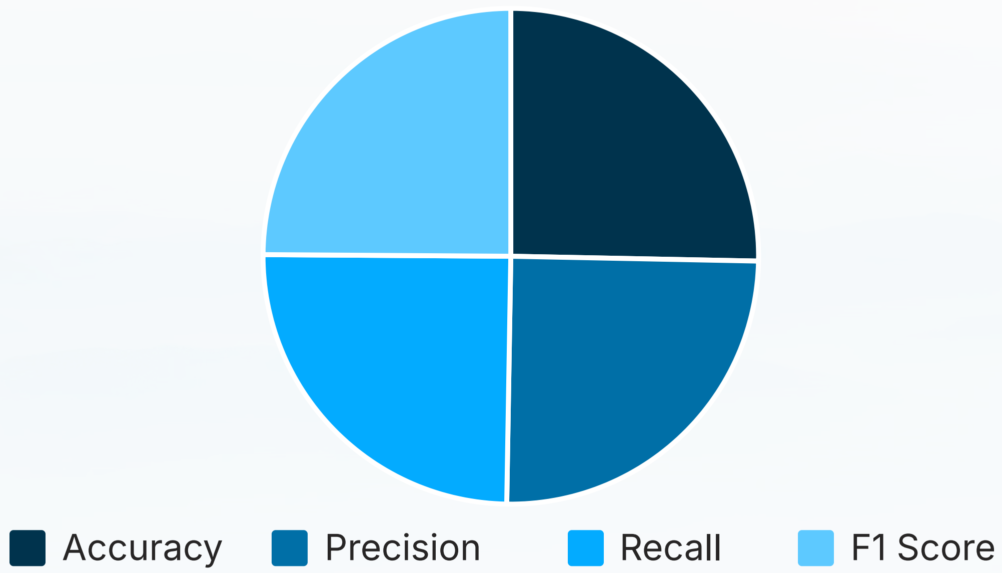
Enter a short news article or title to identify its category.

News Text

"Barcelona beat Real Madrid again"

Sports

Evaluation Results



Performance Highlights

91.5% accuracy on the test set demonstrates strong model performance.

High precision and recall across all four categories indicate balanced predictions.

Confusion matrix analysis showed strongest performance on Sports articles.

Application Interface

User Experience

Simple text input field accepts news headlines or article snippets.

Prediction displays category label and confidence scores.

Interface designed for intuitive use without technical knowledge.

Built with Gradio for rapid deployment

Real-time predictions with minimal latency

Responsive design works across devices

Conclusion & Resources

Project Outcomes

Successfully implemented a complete ML pipeline in NLP

Achieved high accuracy with classical ML approach

Created user-friendly interface for practical application

Resources

All project files available in the GitHub repository:

- Source code and Jupyter notebooks
- Dataset and preprocessing scripts
- Deployment application and documentation
- Comprehensive README with setup instructions